Preprint
UCRL-JC-140288

Data Management Tools

M.N. Ridley and C. Stoker

This article was submitted to the World Water & Environmental Resources Congress Conference, Orlando, Florida, May 20-24

February 12, 2001

U.S. Department of Energy

Lawrence
Livermore
National
Laboratory

Data Management Tools

M. N. Ridley* and C. Stoker**

*Lawrence Livermore National Laboratory, Environmental Restoration Division (ERD), P.O. Box 808, L-528, Livermore, CA 94551-0808; PH (925) 422-3593; FAX (925) 422-2095; email: ridley1@llnl.gov
**Lawrence Livermore National Laboratory, Environmental Restoration Division, P.O. Box 808, L-528, Livermore, CA 94551-0808; PH (925) 422-5725; FAX (925) 422-2095; email: stoker1@llnl.gov

Abstract

What is data management (DM) and why is it important? As described in the *Handbook of Data Management* (Thuraisingham, 1998), data management is the process of understanding the data needs of an organization and making the data available to support the operations of the organization. The ultimate goal of data management is to provide the seamless access and fusion of massive amounts of data, information, and knowledge in a heterogeneous and real-time environment, and to support the functions and decision making processes of an organization. The important questions that need to be asked for proper data management are: who is going to be using the data, what types of data need to be stored, and how will this data be accessed? With these questions answered, the data management system (DMS) can then be created, or an existing system can be modified to meet the needs of the organization.

The real importance of a data management system is to provide the end user with a consistent data set of known quality. The elements of a good data management system should include a system that: is modeled to how the data is collected and processed, is very well documented, has specifically defined data elements, and has supporting data documentation. Supporting documentation includes items like quality control (QC) data that is carried with the analytical data, and meta data (information about the data). Supporting documentation can be anything that is useful to the project and that needs to be accessed with the data.

Data sets get better the more they are used. As errors and inconsistencies are identified and corrected, the data set improves. A good DMS will achieve this because its design promotes use, and the ultimate value of data is in its use rather than its storage. The development and use of Internet access tools, and existing environmental data management systems, can help reduce the effort and cost associated with setting up a DMS. This presentation will discuss the qualities of a good data management system and provide examples.

Introduction

I, Maureen Ridley, would like to provide a little background, so that the reader can understand the direction of this paper and its bias. By trade I am an environmental chemist and I never imagined that I would become an advocate for good data management. I never had any intention of learning about data management when I started reviewing data sets for restoration sites. However, after reviewing a great deal of bad electronic data that had been improperly stored, or not stored at all, I became very interested in how data could be well managed. I have learned a tremendous amount about data

management and DMS thanks to people like Carol Stoker and Patricia Ottesen, who really understand environmental data and how it should be stored and accessed. I personally believe, based on my experience, that one of the greatest wastes of environmental restoration money, is the lack of appropriate data management. Nothing can be more frustrating or time consuming than trying to gather up all of the data for a site. Approximately 90% of my time is spent getting the data and cleaning it up before it is in a useable state. There are many large repository systems presently in use; unfortunately, it is typically very difficult to get data from these systems. When data sets are not used frequently, the data sets are typically of poor quality. Also, there are many people under the impression that data being stored in a GIS system is being well managed. In most cases this is a completely inaccurate assumption, unless a good data management system is feeding data to the GIS system. It is not that a GIS system could not manage the data well, it is that the typical GIS system is not designed for this purpose. Incomplete or inaccurate data from poor or non-functional databases, may result in ill-advised or inappropriate decisions by the technical or management staff.

There are many elements that must be in place to create a good DMS. This paper will try to cover the important aspects of what constitutes a good DMS and what makes it useable. The following topics will be addressed in this paper:

- How a good data management system is created.
- The qualities of a good data management system.
- DMS internet access tools.
- Data conversion issues.

The Creation of a Good Data Management System

As organizations become aware of their need for a DMS, their first reaction is to build a new one to suit their needs. Anytime I hear someone discussing the creation of a new DMS it makes me cringe. The failure rates for the creation of a new DMS are very high. How high, no one really knows, because very few people want to admit that a lot of money was spent on something that failed. Unfortunately, the public only tends to hear about the very expensive failures, such as the California Department of Motor Vehicles system (San Francisco Chronicle, 1995). If data management was so easy, everyone would have a great system and have access to all of their data, anytime they wanted it. How many organizations do you know that have that kind of access?

The requirements for designing a good data management systems can be broken down into 1) the personnel required to accomplish this objective and 2) the appropriate choice of a DMS capable of handling the data needs of the project. Frequently, systems are created by a group of individuals with little understanding of the data that will be stored in the system. It is not enough to just know the elements that need to be stored in a system. To create a truly good system, there needs to be an overlap of understanding between the data users and the computer people. This is the foundation for the creation of a good DMS.

1) The personnel required for creating a good data management system:

- An individual(s) with a complete and thorough understanding of the data and how the data is processed. This is the person(s) who should be trained in data modeling and basic data management

and who may also have relational database development experience. This person or persons should be responsible for designing the model for the system.

- If the individual(s) identified above are not experienced in database development, then it is also necessary to have an individual(s) who has an extensive background in database design with some knowledge and understanding of the data that will be stored in the system.

2) The choice of an appropriate DMS:

- It must be capable of handling the data needs of the program.
  - Identification of the appropriate data elements to cover the project's needs.
  - A good entity diagram or structure with relationships correctly modeled.
  - A flexible data driven system. This is a system that allows for additions and field changes without having to reprogram the entire system when it grows beyond original expectations.
  - Identification of the software and platform necessary to build and successfully use the data management system. This can be either a desktop database management system (such as MS Access or dBase) running on a PC, or a larger SQL compliant relational database management system (RDBMS) running on Unix.
  - Ease of use. Tools that allow the users easy access to the system are more likely to be used.

The Qualities of a Good Data Management System

In the following discussions, Lawrence Livermore National Laboratory's DMS and the enABL Data Management System 2000 (EDMS2000) will frequently be referenced as examples. There are many reasons why these systems are the main examples of this paper. 1) Based on experiences with these systems, both are well-designed systems and well maintained. 2) They have provided reliable service for many years. 3) Neither EDMS or the current LLNL systems has license fees associated with the use of the software. EDMS2000 is the only one of these two systems that is designed to be easily distributed at a minimal cost. The main costs of the EDMS system are the hardware, Oracle® licenses, operation and maintenance of the system. 4) Both systems are well documented. All of the EDMS2000 information is accessible over the internet at http://www.arsenaultlegg.com/. This documentation has made learning about the EDMS2000 system very easy. Most systems available presently do not allow for this kind of up front access. 5) The systems are designed to handle large amounts of data.

A few questions need to be asked before going into the details of what are the qualities of a good DMS.

How Much Data? It is helpful, in the beginning of a project, to know how much data needs to be managed. There are a lot of desktop systems available to manage data and these systems can be good to use as long as there are not a large number of records to be stored and reported on, and it is not a multi-user system. A concern about desktop systems is that there can be a lack of multi-user features such as file management. If just one person is managing data for just one site, then this person's system acts as the main management system. However, if there are several people managing a site and each person has different pieces of data or different versions of the data, this can potentially create incomplete data sets or inconsistent data sets. If the amount of data for a project is

unknown, it is always better to err on the larger side and use a system that is designed to handle larger volumes of data.

What Data Needs to be Managed? Another important question to be answered, is what data should be managed? It is important to ask all of the data users (chemists, engineers, hydrologists, geologists, project managers, etc.), what data they need and how frequently they use it? A typical list of data elements is provided in Appendix A. Appendix A (Environmental Management Electronic Data Deliverable - EMEDD list) is a list of elements that all of the Department of Energy (DOE) sites compiled as a representative elements list. More information and details of the Appendix A list can be found at http://ersmo.inel.gov/edd/edd.html. This type of list can be shorter or much longer depending on the system used. Data quality objectives (EPA 1994) should be used to ensure that the data that is collected and analyzed will be the data necessary to address the projects needs and be of the appropriate quality.

How Much Does Data Management Cost? The cost of data management is very dependent on how much data a project has and how much data tracking is necessary for the project. A rough guideline is to budget 5 to 10% of the restoration budget for data management. This amount may be higher or lower depending on the size of the project. This amount may also sound like a high estimate. This estimate is based on several projects that have excellent data management programs. The question that needs to be asked is: How much is the data worth to you and how useable do you want the data? Nothing is more expensive than unusable data or inaccessible data. One of my favorite quotes about data is from ERD's information system management group leader, Patricia Ottesen, "The most expensive sample ever collected is the one whose results are never used."

A good DMS acts both as a data loading tool and a data management and reporting system for environmental data associated with restoration and monitoring programs. A system should be able to manage analytical laboratory data with information related to sites, locations, soil borings, lithology, well installation and monitoring, soil geotechnical laboratory data, field sampling and field tests, chain-of-custody data, and regulatory requirements. There also needs to be a means to import, translate, and export analytical data of various electronic data deliverable (EDD) formats, thus linking laboratory results to site information and enabling enhanced data reporting capabilities. An electronic project archive of known quality, with historical data that are easily accessible by different parties for use in future environmental projects, is a key element of a good system.

A good system is built on a relational database, such as Oracle® or Sybase® for the PC, with screens developed using a program like Oracle® Developer. The DMS is a data driven system. This allows for optimum flexibility so that the data elements may be added or adjusted to meet a project's specific needs. Such potential changes to the system should encompass all major components of the software including entry screens, import/export functions, security/revision tracking, and consistency checking routines. This type of flexibility makes it possible to easily migrate or add information to an existing system. Hence if an organization wished to keep an existing system's design and content, the data from the original system could be moved into a data driven system without translating, reorganizing, or renaming the original data.

The following sections discuss in greater detail important features of a good DMS, such as Entry Screens, Import, Export, Consistency Checking (CC), Administration, and Reports.

Entry Screens**.** Data entry screens allow manual input of geotechnical, laboratory, field, groundwater measurement, and lithological data, as well as project management and regulatory information. The entry screens provide a means of interactively viewing and modifying the DMS. These screens facilitate consistent data entry by checking values for logical entry (e.g., the beginning depth is less than the ending depth). Screen formats may be modified and additional screens may be added dynamically by the user. The system's modifiability and dynamic tasking accommodates user preferences, particularly organization-specific information and business practices.

Import**.** The import function consists of screens to locate and configure import files. Import validation procedures need to be provided to ensure that the imported information is structurally sound. Users need to be alerted to structural inconsistencies via some type of on-line error report generated by a consistency checker. A good DMS allows for the import of various electronic deliverable formats into the same DMS. At present, it seems that every organization has his or her own electronic data deliverable and no one wants to change. So, instead of asking everyone to agree on one data deliverable, which by the way will never happen, have a DMS that can handle many types of electronic data formats. EDMS2000 allows for the import of the EDF, IRDMIS, ERPIMS, NAVY, DEEMS, and the EMEDD formats.

Export**.** The export function allows the generation of reports and export files based on user-defined criteria. Also, set up functions utilize screens that allow selection of restrictive criteria, database query routines, and destination selection (i.e., interactive browse, reports, and data exports).

Administration (Security/Revision Tracking). The security/revision tracking functions allow restricted access accounts and tracking of data modification. The DMS administrator has the ability to restrict user rights to functions such as read-only, data modification, report generation, and import functions. A good system tracks all revisions and is able to generate reports showing data modifications of imported and manually entered data, the user modifying the data, and the date and time that it was modified. The revision tracking information needs to be queried and reported in the same fashion as the other environmental data within the DMS.

Consistency Checking**.** The consistency checking locates files, selects tables for review, and initiates electronic data checking. This procedure is to ensure that the data are structurally sound before being placed in the DMS. A user should be alerted to structural inconsistencies via an error report.

Web tools (reports)

The elements in the previous sections are the foundation for a good DMS; however, it is the internet access that creates a useful system for the data users. LLNL's Environmental Restoration Program has been using Internet access tools (web tools) to access their DMS since 1995. This system has allowed technical people, regulators and the community to access the LLNL's site data. Another DMS, the EDMS2000 has also added internet access web tools to their system.

Web tools and wizards assist with the management and dynamic reporting of environmental data that can be accessed by multiple approved users from anywhere in the world. These reports are accessed through the web tools via either Internet Explorer® or Netscape®. Reports are generated

using the web tools that have been added to more fully utilize the data contained within a DMS. Users require no special training and are guided by help tools commonly referred to as wizards. Web tools are built to retrieve specific types of data. A tool can be built to retrieve any type of data that a user needs, such as summary data reports and statistical analysis reports. These tools can also provide visualization of data, such as graphs and contour plots. The following are descriptions of just a few tools that are attached to the LLNL's DMS and to EDMS 2000. LLNL presently has over 30 different web tools to suit their needs and EDMS 2000 has over 15 tools. Tools are very easy to create once the initial system is in place, thus the number of tools quickly grows. To obtain a first hand view of these types of internet access tools, a demonstration web site has been setup at http://www.arsenaultlegg.com/software.htm.

GIS Contouring**.** Topographical map information may be imported to enable GIS plots where such maps are available. The GIS contour plots can be generated utilizing specified base maps and underlying data stored in the DMS. The generated plots can be zoomed for detailed viewing.

Cost-Effective Sampling**.** This Web tool enables cost-effective sampling (CES) analyses and reporting. The CES statistical tool utilizes data residing in the DMS to provide frequency recommendations for routine monitoring programs. The system allows Real-Time adjustment of constants and data ranges. Checking will occur to verify the statistical viability of the data. Wizards are provided to assist in use of this tool and the creation of CES reports that may be generated from the data session.

Time Series Graphs**.** This Web tool enables time series data to be viewed in graph form. Users may enter a time interval, specifying start and end times, in order to easily see the changes of a parameter over time. The Time Series Graphs can display up to four different customized data series in line graph format. These reports are generated from data residing in the DMS that can be viewed in table format or downloaded to a Microsoft® Excel spreadsheet.

Data Extraction/Reduction Reports**.** Data Reduction Reports can be generated to display specified data. The selected data can also be downloaded to an Excel spreadsheet for further format manipulation using the tools of Excel.

Project Management Reports**.** Wizards guide users in the creation of pertinent project management reports that determine site status, evaluate project schedules, and track regulatory requirements.

Translation

Translation and cleanup of a data set must be done before the data can be put into a DMS. If you are just starting a project, then the electronic data can be collected from the field and analytical laboratory from the beginning and a translation is not necessary. However, if you have hard copy or electronic data that needs to be entered in a DMS, then this section maybe very helpful. The following discussion is an example of a translation and cleanup of a data set. During the course of a translation of the Riverbank Army Ammunition Plant (RBAAP) site data, from Installation

Restoration Data Management Information System (IRDMIS) to EDMS2000, several difficulties related to the IRDMIS data set were encountered.

The difficulties vary in terms of the actual impact to the translating effort. To translate data from one system to another, the structures of the systems need be known. In the translation of the IRDMIS data, no information was available on the database structure. The missing information included both the data relations (how one group of data links to another) and primary key identification (what makes a record of information unique within a data group). Without this information, any interpretation of the data within the IRDMIS system is, at best, an educated guess and that was exactly how the IRDMIS data had to be handled.

DMS Documentation A data dictionary is a document that is also necessary for the translation of data. A data dictionary contains a list of all files in the database, the number of records in each file, and the names and types of each field. Data dictionaries do not contain any actual data from the database, only bookkeeping information for managing it. In the IRDMIS data example, the IRDMIS Data Dictionary was used to help in the translation. While the IRDMIS documentation was fairly complete in its description of the informational fields and coded values, its discussion of the data relations was incomplete in terms of describing what makes a record of information unique (primary key), as well as describing how the groups of data link together (data relations). What makes the data records unique and how data groups are linked, are key to a user's understanding and interpretation of the data stored within the system. For example, if a geospatial location has multiple location names, how does one location name connect to another to pool data to a single geospatial point? Because the system's structure documentation is the basis for any database implementation, the lack of this documentation is a significant barrier to data interpretation.

Specific Data Issues Many different data issues may be encountered during a translation; however this discussion will review the most common data problems.

- Informational fields documented as "required" by the data system, do not contain data. (Some of these informational fields are basic to data interpretation such as matrix, elevation, and units of measurement).
- Some of the coded values within the data (which are used to represent information such as analytical method) are not documented in the data dictionary.
- Entries that are supposed to be limited to specific data tables, per the documentation, appear in tables in which they do not belong.
- Locations are identified by more than one name. Because the connections between the data groups are not documented, if a location is named one thing in one data group and something different in the next data group, the pieces of information may not link via location, as would be standard in a geospatial system. This significantly hinders the interpretation of the data.

Translation Conclusions There is a key distinction that needs to be made concerning data translation and data scrubbing/grooming. The data translation is the process by which the data is moved from one DMS to another, such as moving data from IRDMIS to EDMS 2000. The data scrubbing/grooming is the process by which the data is examined and cleaned up to make the data set useable. As in the example of the initial Riverbank data translation and data scrubbing/grooming, both were very labor intensive processes. Subsequent translations of IRDMIS

data will be much easier due to the automated translation tool that was developed during the Riverbank data translation. However, the data scrubbing/grooming will be just as labor intensive each time data from a new site is examined and cleaned up. Unfortunately, data systems that act solely as an archive record, with little or no retrieval activity, tend to contain data with many internal inconsistencies and perplexing ambiguities.

Conclusions

The use of a good data management system can allow data to be used to its maximum potential. The overall process of translating the data in to a DMS and cleaning up the data may sound very daunting, however the rewards can be immense. What the web tools/DMS can ultimately mean to a project is:

- Fast, efficient and reliable user access and visualization of a data set of known quality.
- Optimization of the cleanup and faster closure of a site.

Since RBAAP has had access to their data through the web tools/DMS, RBAAP has cut their treatment process operation by 50% and are in the process of reducing their long term monitoring cost by 40%. The most important return on investment for data is in their use rather than their storage.

This work was performed under the auspices of the U.S. Department of Energy by University of California Lawrence Livermore National Laboratory under contract No. W-7405-Eng-48.

References:

EPA, 1994, *Guidance for the Data Quality Objective Process:* EPA QA/G-4, U.S. Environmental Protection Agency, Washington, DC.

San Francisco Chronicle, *Boss of DMV Resign – Dispute on Reform Efforts*, October 11, 1995.

Thuraisingham, B., 1998, *Handbook of Data Management*, Auerbach, New York, NY.

Appendix A:

EMEDD List of Elements

Analyzed	Detection Limit Type	Matrix ID	QC Precision Limit Type
Batch Date	Dilution Factor	Original Lab Analysis ID	Quantitation Limit
Batch ID	EDD ID	Original Lab Sample ID	Quantitation Limit Type
Batch Procedure ID	EDD Version	Pay Item	Reporting Limit
Batch Procedure Name	Filtered	Percent Moisture	Reporting Limit Type
Batch Type	Instrument ID	Percent Recovery	Result
Client Method ID	Lab Analysis ID	Percent Recovery Limit High	Result Basis
Client Qualifiers	Lab ID	Percent Recovery Limit Low	Result Text
Client Sample ID	Lab Matrix ID	Percent Recovery Limit Type	Result Units
Collected	Lab Name	Percent Solids	Retention Time
Comment	Lab Procedure ID	pH	Sample Type
Counting Error	Lab Procedure Name	Preservative	Temperature
Counting Error Type	Lab Qualifiers	QC Linkage	Uncertainty
Custody ID	Lab Receipt	QC Precision	Uncertainty Type
Date Submitted	Lab Reporting Batch	QC Precision Type	
Detection Limit	Lab Sample ID	QC Precision Limit High	